STAT 651 Chapter 5.1–5.5

Zepu Zhang
December 16, 2010

1 Concept of random (iid) sample

Definition 5.1.1, p. 207. Random sample or iid random variables.

Note 1. Two ways to call the same thing: random samples from a certain population or iid random variables with a certain distribution.

2. The "random sample" are a bunch of random variables rather than a dataset. When each X_i has a value x_i, we have a particular set of observations. I think it is also common to see statements like "$x_1, \ldots, x_n$ are an iid sample from distribution...

3. $X_1, \ldots, X_n$ form a random vector. We can easily write the joint pdf (or pmf) of this random vector because they are iid, and this joint pdf is an important tool.

4. We often provide information about the common distribution through X_1, e.g. $E(X_1) = 3$, $\text{var}(X_1) = 4$, etc. This is information for all the X's, not only X_1.

5. If the common distribution has parameters, these parameters are also parameters for the distribution of the sample, of course. If the parameters are unknown, then a very common task to to estimate the parameters based on the sample. To this end, one would come up with an estimator for the parameter(s); the estimator is a function of the sample. The value of the estimator based on an observation of the sample (one value from each X_i) is an estimate. Then one would usually study properties of the estimator such as unbiasedness, variance, etc.

6. A maximum likelihood estimator of parameter θ is the θ that maximizes the joint pdf (or pmf) of the sample. This is a fundamental way of finding an estimator.

2 Statistics and sampling distribution

Definition 5.2.1, p. 211. Statistic and sampling distribution. A (real values or vector valued) function of a sample is called a statistic.

The distribution of a statistic is called its sampling distribution.

Note Typical, useful statistics: sample mean, sample variance, order statistics (the smallest of $X_1, \ldots, X_n$; the second smallest of $X_1, \ldots, X_n$;...; the largest of $X_1, \ldots, X_n$).

3 Sample mean and sample variance

Definition 5.2.2., p. 212. Sample mean.

Definition 5.2.3., p. 212. Sample variance.

Note Notation: $\overline{X}$, S^2, S, $\overline{x}$, s^2, s.

Theorem 5.2.4, p. 212.

Proof.

Compare with analogous results for a population.

Theorem 5.2.5, p. 213.

Proof. Proof of (5.2.2) can use generalized theorem 4.5.6 (p. 171) (in lecture notes) and theorem 4.6.12, p. 184.

4 Sampling distribution of the sum of random variables

Theorem 5.2.6, p. 213.

Note 1. Condition of this theorem: μ and σ^2 exist (ie finite).

2. $\overline{X}$ and S^2 are unbiased estimators of μ and σ^2.

3. Sample mean has a smaller spread than the original variable.

Find the distribution of sample means

This is not always possible. But at least we know about the mean and variance of a sample mean, according to theorem 5.2.6.

Two ways mentioned in the book: theorems 5.2.7 and 5.2.9.

Theorem 5.2.7, p. 215.

Example 5.2.8. (Application of theorem 5.2.7.) The result is important! The proof is also simple enough.

Theorem 5.2.9, p. 215.

Example 5.2.10, p. 216. (Application of theorem 5.2.9.) This example once again shows that Cauchy has a heavy tail. In fact, the sample mean has as much uncertainty as X itself. Note theorem 5.2.6 does not apply to Cauchy. Although Cauchy is "centered" at and symmetric about 0, we do not call 0 its "mean". Cauchy has no finite moments.

(Skip the material in section 5.2 after example 5.2.10.)

5 Sample mean and sample variance of normal

Theorem 5.3.1, page 218.

Note 1. All three results (a, b, and c) are important. Perhaps result a is not often used directly.

2. Result b is proven in example 5.2.8, p. 215, via mgf.

3. Proof for results a and c is somewhat involved. You are not required to do proofs of this sophistication in exercises.

Lemma 5.3.2, page 219. Facts about χ^2.

Note These results are quite basic and elegant. You should be familiar with them. Result b has a special case: the sum of squares of n iid standard normal variables is χ^2_n.

Lemma 5.3.3, page 220. Covariance and independence for normal.

Note 1. The statement of this lemma is quite tedious. Don't be confused by it. We'll approach this statement through the following matrix notation.

2. The main point of this lemma is the following. We know independence suggests $\text{cov}(x, y) = 0$, but the converse is not true. If $\text{cov}(x, y) = 0$, X and Y may or may not be independent. However, if X and Y have a joint normal distribution (we say they are "jointly normal"), then the converse is true.

3. The proof is not required.

Suppose we have a normal vector, $\vec{X} = [X_1, \ldots, X_n]'$, and a $(k + m) \times n$, where $k + m \leq n$, matrix $\boldsymbol{T}$, then

$$\vec{Y} = \boldsymbol{T}\vec{X}$$

is a $(k + m)$-dimensional normal vector. The condition $k + m \leq n$ is there because we can't get the distribution of more than n variables given the joint distribution of n variables.

$$\vec{Y} \sim N\big(\boldsymbol{T}E(\vec{X}),\, \boldsymbol{T}\,\text{cov}(\vec{X})\boldsymbol{T}'\big)$$

If we take the first k rows of $\boldsymbol{T}$ to be $\boldsymbol{A}$ and the last m rows to be $\boldsymbol{B}$, then $\vec{Y}$ is composed of the first k elements $\vec{U} = \boldsymbol{A}\vec{X}$ and the last m elements $\vec{V} = \boldsymbol{B}\vec{X}$. $\vec{U}$ and $\vec{V}$ are two random vectors that are jointly normal. The mean of $\vec{Y}$ can be partitions into a part for $\vec{U}$ and a part for $\vec{V}$. The cov matrix of $\vec{Y}$ can be partitions into 4 blocks, for $\text{cov}(\vec{U}, \vec{U})$, $\text{cov}(\vec{U}, \vec{V})$, $\text{cov}(\vec{V}, \vec{U})$, and $\text{cov}(\vec{V}, \vec{V})$. The two "self" blocks are both symmetric. The two "cross" blocks are transpose of each other. The vectors $\vec{U}$ and $\vec{V}$ are independent if and only if the two "cross" covariances are 0 (that is, all elements are 0).

This statement certainly can be used when both $\vec{U}$ and $\vec{V}$ are univariate.

6 t and F distributions

No need to memorize the density formulas for them. Should know how they occur:

1. $$\frac{\text{standard normal}}{\sqrt{\chi^2_p/p}} \sim t_p$$

   Note the standard normal and the χ^2 should be independent. Also note how theorem 5.3.1c is used for (5.3.5).

   We say the pattern on the LHS is t. Of course a t variable does not have to arise that way—a rv has a t distribution as long as its distribution conforms to the t formula. But the above is a common mechanism by which a t variable happens. The most common pattern is (5.3.4), p. 222.

2. $$\frac{\chi^2_p/p}{\chi^2_q/q} \sim F_{p,q}$$

   Note that two χ^2's should be independent.

   About how such variables occur: similar comments, see above about t.

   F distribution is often used in testing whether two variances are equal. Note theorem 5.3.1c, p. 218, connects sample variance with χ^2.

Additional note:

1. You should know the relation between standard normal and t: shape, symmetry, tail fat or thin, spread. When $p \to \infty$, t approaches standard normal.

2. t_1 is Cauchy.

3. In addition to tests, t is often useful when we need a model that has the nice normal bell shape and symmetry but has heavier tails than normal so that extreme values have a larger chance to occur.

7 Order statistics

Regarding "order statistics", you should know the concept and can work on the two simplest cases, $X_{(1)}$ (the smallest) and $X_{(n)}$ (the biggest).

Definition 5.4.1, p. 226.

Note 1. Each "order statistic" is a "statistic', i.e. a function of a sample.
2. The notation $X_{(1)}$ is not universal. Just follow the definition of symbols in a local context.

Some other statistics that are defined based on order statistics:
sample range
sample median
sample quartiles, lower quartile, upper quartile
inter-quartile range
sample percentiles (p. 227, below definition 5.4.2)
trimmed mean (e.g. throw out the largest and smallest, average the rest)

Note 1. "Median" and "quartiles" are special "percentiles"; "median" is a special "quartiles".
2. These are found after putting the sample in order. But these are not "order statistics", because they may not be any of the sample members directly—we often need to interpolate two adjacent order statistics to get a certain percentile.
3. How to interpolate: as far as I have seen it's not necessarily taken to be the middle point (the mean) between two adjacent order statistics. Just follow the author's definition in any particular study.

I re-checked definition 5.1.1 and thought this is worth emphasizing: $X_1, \ldots, X_n$ is called a random sample of size n, NOT n random samples. (I don't think this is crucially important and strict, but it's good to make a clear choice of terminology

and stick to it.)

The largest order statistic, $X_{(n)}$

(X is continuous.)

$$F_{X_{(n)}}(x) = P(X_{(n)} \le x) = P(X_i \le x, i = 1, \ldots, n) = \big[P(X_{(1)} \le x)\big]^n = \big[F_X(x)\big]^n$$

$$f_{X_{(n)}}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\big[F_{X_{(n)}}(x)\big]^n = n\big[F_X(x)\big]^{n-1} f_X(x)$$

The smallest order statistic, $X_{(1)}$

(X is continuous.)

$$F_{X_{(1)}}(x) = P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) = 1 - P(X_i > x, i = 1, \ldots, n) = 1 - \big[1 - F_X(x)\big]^n$$

$$f_{X_{(1)}}(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{X_{(1)}}(x) = n\big[1 - F_X(x)\big]^{n-1} f_X(x)$$

Example $X \sim \text{unif}(0, \theta)$, $\theta > 0$.

Noticing

$$f(x) = \begin{cases} 1/\theta, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}; \quad F(x) = \begin{cases} 0, & x \le 0 \\ x/\theta, & 0 < x < \theta \\ 1, & x > \theta \end{cases}$$

we get

$$f_{X_{(n)}}(x) = n(x/\theta)^{n-1}(1/\theta) = nx^{n-1}\theta^{-n}, \quad 0 < x < \theta.$$

$$f_{X(1)}(x) = n(1 - x/\theta)^{n-1}(1/\theta) = n(\theta - x)^{n-1}\theta^{-n}, \quad 0 < x < \theta.$$

Example Independent, but not identically distributed continuous random variables. Consider independent random variables X_1 and X_2 where their respective pdf's are given by $f_1(x) = \frac{1}{3}x^2 I(-1 < x < 2)$ and $f_2(x) = \frac{5}{33}x^4 I(-1 < x < 2)$.

One has the distribution functions $F_1(x) = \int_{-1}^{x} \frac{1}{3}t^2 \,\mathrm{d}t = (x^3 + 1)/9$ and $F_2(x) = \int_{-1}^{x} \frac{5}{33}t^4 \,\mathrm{d}t = (x^5 + 1)/33$. Hence, for $-1 < x < 2$, the distribution function of $X_{(2)}$, the larger order statistic, can be found as follows:

$$F_{X_{(2)}}(x) = P(X_1 \le x, X_2 \le x) = F_1(x)F_2(x) = \frac{1}{297}(x^3+1)(x^5+1)$$

Differentiate $F_{X_{(2)}}(x)$ to get the pdf.

The same idea easily extends for n independent, but not identically distributed continuous random variables.

8 Two types of stochastic convergence

Stochastic convergence is a mathematical concept intended to formalize the idea that a sequence of essentially random or unpredictable events sometimes is expected to settle into a pattern.

These concepts are defined in terms of a series of random variables, $X_1, X_2, \ldots, X_n$. The behavior of these r.v. changes as $n \nearrow$. We consider two patterns as $n \to \infty$:
(1) X_n stays arbitrarily close to a constant value or the value of a "target" r.v. with very high probability;
(2) The distribution of X_n gets increasingly similar to a certain distribution.

Definition 5.5.1, p 232. "in probability"

$$P\big(|X_n - X| > \epsilon\big) \xrightarrow{n\to\infty} 0, \quad \forall \epsilon > 0$$

Notation: $\xrightarrow{P}$.

A common special case is that X is a constant, say θ. In fact, we can understand the concept for this special case, then $X_n \xrightarrow{P} X$ is equivalent to $X_n - X \xrightarrow{P} 0$.

Definition 5.5.10, p 235. "in distribution".

$$P(X_n \le x) \xrightarrow{n\to\infty} P(X \le x),$$

at every continuity point of $F_X(x)$.

A special case is that X is a constant.

Also called "in law" or "weak convergence".

Notation: $\xrightarrow{D}$ or $\xrightarrow{L}$.

Note 1. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$, that is "in probability" is stronger than "in distribution". This is theorem 5.5.12, p. 236.

2. If X is a constant, say θ, then these two types of convergence are equivalent, that is, $X_n \xrightarrow{P} \theta \iff X_n \xrightarrow{D} \theta$. This is theorem 5.5.13, p. 236.

Example 5.5.11, p. 235. Let $X_1, \ldots, X_n$ be iid U$(0,1)$. Then $X_n \xrightarrow{P} 1$ and $n(1 - X_n) \xrightarrow{D} Q \sim \exp(1)$.

9 Two basic theorems

Theorem 5.5.2, p. 232. WLLN. iid, $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2 < \infty$,

then

$$\overline{X}_n \xrightarrow{P} \mu$$

Theorem 5.5.15, p. 238. CLT. iid, $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2 < \infty$, then

$$\frac{\sqrt{n}\big(\overline{X}_n - \mu\big)}{\sigma} \xrightarrow{D} Z \sim N(0,1)$$

This can also be written as

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)$$

or

$$\sqrt{n}\big(\overline{X}_n - \mu\big) \xrightarrow{D} \sigma Z \sim N(0,\sigma^2)$$

or

$$\overline{X}_n - \mu \xrightarrow{D} \frac{\sigma}{\sqrt{n}} Z \sim N(0,\sigma^2/n)$$

or

$$\overline{X}_n \xrightarrow{D} \mu + \frac{\sigma}{\sqrt{n}} Z \sim N(\mu,\sigma^2/n)$$

That these forms are equivalent is not all obvious. The Slutsky's theorem (later) guarantees this.

Note 1. In both theorems, the series is the sample mean as the sample size increases. WLLN says the sample mean approaches the population mean; CLT says the standardized sample mean approaches a standard normal r.v.

2. Note we do not require X_i to be normal. The conditions of the theorems are rather mild (usually met), hence very general. If X_i is normal, we have learned that $\overline{X}_n$ is normal whatever the sample size (the distribution is exact; no convergence or approximation involved).

Example 5.5.16, p. 239.

10 Getting new "limits" from known ones

WLLN, CLT, and the following theorems help us get "new" limits from known ones.

Theorem 5.5.4, p. 233. Continuous mapping theorem. Let $g : \mathcal{R} \to \mathcal{R}$ be a continuous function. Then
(a) $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$
(b) $Y_n \xrightarrow{D} Y \Rightarrow g(Y_n) \xrightarrow{D} g(Y)$

Example $\overline{X}_n \xrightarrow{P} \mu \Rightarrow \overline{X}_n - \mu \xrightarrow{P} 0$.
$S^2 \xrightarrow{P} \sigma^2 \Rightarrow S \xrightarrow{p} \sigma$.

Example From CLT we have $\sqrt{n}\big(\overline{X}_n - \mu\big)/\sigma \xrightarrow{D} N(0,1)$. Then we immediately get

$$n\big(\overline{X}_n - \mu\big)^2/\sigma^2 \xrightarrow{D} Z^2 \sim \chi^2_1$$

since the square of a standard normal is χ^2_1. Compare this with a result we have about S^2.

Theorem 5.5.17, p. 23. Slutsky's theorem. If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} a$ (a constant; P or D are the same). Then

(a) $X_n \pm Y_n \xrightarrow{D} X \pm a$.

(b) $Y_n X_n \xrightarrow{D} aX$.

(c) $X_n/Y_n \xrightarrow{D} X/a$ provided that $P(Y_n = 0) = 0$ for all n and $v \neq 0$.

Example 5.5.18, p. 240.

If we know the distribution of X, then we may get the distribution of $g(X)$ by transformation, if the function g is nice. Although the distribution of $g(X)$ may be written out this way, it may not be easy to further analyze it.

Example Suppose we have $X_n \xrightarrow{D} Z \sim N(0,1)$. Then by the continuous mapping theorem we know $X_n^3 \xrightarrow{D} Z^3$, the cube of a standard normal. But what does the distribution of Z^3 look like? We still don't know.

Let's aim for a lower target. What are the mean and variance of $g(X)$? This is not easily known unless g is a linear function. In general, $Eg(X) \neq g\big(E(X)\big)$ (remember this)!

Example $E(S^2) = \sigma^2$ does NOT suggest $E(S) = E(\sqrt{S^2}) = \sqrt{E(S^2)} = \sigma$. The square-root is not a linear function.

But we can get approximations to $Eg(X)$ and $\operatorname{var} g(X)$. Before getting to that, a brief review of Taylor series. You need to know the following about Taylor series.

Theorem Taylor series expansion:

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots$$

Note 1. Note a rather common notation for higher-order derivatives: $f^{(i)}$. If only the first two are needed, then it's also common to use f' and f''.
2. Necessarily we'll discard some later terms and accept an approximation. Of course, the closer x is to a, the better the approximation.

3. Apparently we are assuming the derivatives that are written here exist.
4. If higher-order derivatives disappear (are 0), say the 5th and above, then those later terms vanish. It's not a problem and we're not losing anything—the Taylor series is still there, only that later terms are 0.
5. People often take just the first-order approximation because it's easy to work with, and sometimes pretty good already.

Exercise In what situations a first-order approximation is pretty good? (Answer: $|x-a|$ is small and/or $|f''(a)| \ll |f'(a)|$, meaning f is almost linear in the neighborhood of a.)

6. The series can be written concisely as $\sum_i^\infty \frac{f^{(i)}(a)}{i!}(x-a)^i$. Note $f^{(0)}$ means f.

Theorem (Informally) Write

$$f(x) = \sum_{i=1}^{r} \frac{f^{(r)}(a)}{i!}(x-a)^i + \text{remainder}$$

If we take the first r terms as an approximation, then the remainder (i.e. error in the approx) decreases faster than the highest-order term (the rth term) used in the approximation. A more formal way to say this is theorem 5.5.21, p. 241.

Theorem (Informally) It is often convenient for analysis to write the remainder as $\frac{f^{(r+1)}(a^*)}{(r+1)!}(x-a)^{r+1}$, where a^* is some value between x and a.

Now let's take a first-order approx:

$$g(X) \approx g(\theta) + g'(\theta)(X-\theta)$$

where $\theta = E(X)$. Then we see

$$\begin{aligned} E\big(g(X)\big) &\approx g(\theta) + g'(\theta)\big(E(X)-\theta\big) = g(\theta) \\ \operatorname{var}\big(g(X)\big) &\approx \big(g'(\theta)\big)^2 \operatorname{var} X \end{aligned}$$

where $g'(\theta)$ is the derivative of the function g evaluated at $E(X)$.

Now the relation between $g(X)$ and $g(\theta)$ is somewhat like that between $\overline{X}$ and μ. We know the distribution of $\overline{X}$ approaches normal when $\operatorname{var}(\overline{X})$ keep decreasing in a series. Now if the variance of $g(X)$ keeps decreasing, that is, $\operatorname{var}(X)$ keeps decreasing, then we have a similar result.

Theorem 5.5.24, p. 243. Delta method, or Mann-Wald theorem. Let

$X_n \xrightarrow{D} N(\theta, \sigma^2/n)$, and suppose $g'(\theta)$ exists and is not 0, then

$$g(X_n) \xrightarrow{D} N\Big(g(\theta),\ \big[g'(\theta)\big]^2\sigma^2/n\Big)$$

(I think this form is more natural to remember than the form in the book.)

Note $\text{var}(X_n) \to 0$, i.e. $X_n \xrightarrow{P} \theta$. This is key. If X_n is just normal (with a stable variance), we wouldn't know the distribution of an arbitrary function, $g(X_n)$, of it.

Proof 1 Let's prove the "standard" form

$$\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{D} N\big(0,\ (g'(\theta))^2\sigma^2\big)$$

Let $U_n = \sqrt{n}(X_n - \theta)$ and $V_n = \big(g(X_n) - g(\theta)\big)/(X_n - \theta)$. Then the lhs is U_nV_n.

From $U_n \xrightarrow{D} N(0, \sigma^2)$ as $n \to \infty$, by Slutsky's theorem we conclude $X_n - \theta = \frac{1}{\sqrt{n}}U_n \xrightarrow{D} 0 \cdot N(0, \sigma^2) = 0$. Thus, $V_n \xrightarrow{P} g'(\theta)$ by the definition of $g'(\theta)$. Using Slutsky again,

$$\text{lhs} = U_nV_n \xrightarrow{D} g'(\theta)U_n \sim N\big(0,\ (g'(\theta))^2\sigma^2\big)$$

Proof 2 By Taylor's theorem,

$g(X_n) = g(\theta) + g'(\theta^*)(X_n - \theta)$ for some θ^* between X_n and θ.

From $X_n \xrightarrow{P} \theta$ we get $\theta^* \xrightarrow{P} \theta$, hence $g'(\theta^*) \xrightarrow{P} g(\theta)$. Then by Skutsky,

$$\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{D} \sqrt{n}g'(\theta)(X_n - \theta) \xrightarrow{D} g'(\theta) \cdot N(0, \sigma^2) \sim N\big(0,\ (g'(\theta))'\sigma^2\big)$$

Example Suppose $\sqrt{n}\big(\overline{X}_n - \mu\big) \xrightarrow{D} N(0, 1)$. Find the distributions of $\sqrt{n}\big(\overline{X}_n^2 - \mu^2\big)$ and $\sqrt{n}\big(\sqrt{\overline{X}^2} - \sqrt{\mu}\big)$ without using Mann-Wald.

Notice

$$\sqrt{n}\big(\overline{X}_n^2 - \mu^2\big) = \big\{\sqrt{n}\big(\overline{X}_n - \mu\big)/\sigma\big\}\big\{\sigma\big(\overline{X}_n + \mu\big)\big\}$$

Denote the two terms by U_n and V_n. We have $U_n \xrightarrow{D} N(0, 1)$ and $V_n \xrightarrow{P} 2\mu\sigma$. By Slutsky,

$$\text{lhs} = U_nV_n \xrightarrow{D} 2\mu\sigma \cdot N(0, 1) \sim N(0,\ 4\mu^2\sigma^2)$$

(Note the notation mixing variables and distributions is not very good. Introducing a variable $Z \sim N(0, 1)$ can help.)

Note Using Mann-Wald, we can do more generally:
$\sqrt{n}\left(\overline{X}_n^q - \mu^q\right) \xrightarrow{D} ....$

Example Let $X_1, \ldots, X_n$ be iid Poisson(λ) with $\lambda > 0$. Find the distribution of $\sqrt{n}\left(\overline{X}_n^3 - \lambda^3\right)$.

Example 5.5.23, p. 242.

Example 5.5.25, p. 243.